Universidade de Brasília
Departamento de Estatística

Deficiências da estatística Kappa na concordância entre avaliadores e medidas alternativas

por
Hudson Thiago Afonso dos Santos

Brasília
2015

HUDSON THIAGO AFONSO DOS SANTOS - RA:10/00054447

Deficiências da Estatística Kappa na concordância entre avaliadores e medidas alternativas

Proposta Final de Estágio supervisionado obrigatório apresentado à Universidade de Brasília como requisito parcial à obtenção do título de bacharel em Estatística.

Orientadora: Prof. Cibele Queiroz da Silva, Ph.D.

Brasília
2015

Ao poderoso Deus e a toda minha família, aos meus
amados pais Bráulio e Jaqueline e meu irmão Halisson
que sempre acreditaram, sonharam e viveram comigo cada momento.

Hudson Thiago Afonso dos Santos

Agradecimentos

Minha gratidão é em primeiro lugar ao Senhor Deus que sustentou-me, carregando-me nos braços em todos os momentos da minha graduação, sendo-me sempre fiel na minha inconstância.

Agradeço à minha mamãe, Dona Jaqueline, que sempre esteve, está e estará ao meu lado em todos os momentos; aquela que acreditou em mim quando porventura perdi minhas forças e me deu de volta o ânimo, seja com um sorriso, um lanchinho preparado com amor, ou com puxões-de-orelha e um empurrãozinho quando precisei.

Ao meu papai, Seu Bráulio, que fez de tudo para que eu 'chegasse lá'! Por cada cuidado, seja ao me buscar na parada em dias chuvosos ou ao adormecer apenas quando eu entrasse em casa, meus sinceros agradecimentos!

À minha orientadora, professora Cibele, que me acolheu e creio que tirou o melhor de mim e teve paciência para comigo nas minhas demoras, meu muito obrigado!

A todos que de alguma forma me impulsionaram a chegar onde cheguei, seja nas longas e agradáveis conversas nos corredores da UnB, nas manhãs, tardes e noites de estudo. Aos meus colegas de curso que me animavam, ensinavam e divertiram muito ao longo desses anos! Sem vocês, eu não conseguiria! Sou muito grato por todos que Deus colocou na minha vida para que eu aprendesse e desse valor ao preço do meu sonho. Através da presença de cada um comigo, tenho um coração agradecido a Deus por sua misericórdia na minha vida. Obrigado a todos!

Resumo

Avaliar a intensidade da concordância entre dois ou vários avaliadores é muito comum nas ciências sociais, comportamentais e médicas. Essa Monografia consiste, então, em versar sobre índices de concordância entre avaliadores dando uma ênfase especial à estatística Kappa, comumente utilizada como coeficiente de mensuração da intensidade de concordância inter-avaliadores. Fornecendo insumos para as discussões sobre as deficiências deste índice, aponta-se limitações para sua prática, implementa-se computacionalmente métodos sugeridos por Gwet (2002) e indica-se a adoção, para alguns casos, de um coeficiente mais robusto: o índice AC1.

Palavras-chave: Concordância entre avaliadores, índice Kappa, índice AC1

Índice

1 Introdução

A estatística é utilizada por muitos pesquisadores para analisar e entender um conjunto de dados relevantes ao seu estudo particular. Estes necessitam extrair informações dos seus dados para compará-los com outros resultados e para julgar sua adequação a alguma teoria.

Dessa forma, a Estatística é uma ciência que dá suporte para planejamento de experimentos, estudos e métodos de sumarização de dados, apresentação desses mesmos dados, interpretações e análises. Logo, a Estatística auxilia na tomada de decisão por estimar e quantificar o quanto de incerteza temos com respeito aos valores dos parâmetros desconhecidos no modelo estatístico que descreve um dado fenômeno.Bussab and Morettin (2010)

Um problema estatístico bastante frequente é o da Análise de Concordância entre Avaliadores. Considere que um conjunto de n indivíduos será analisado por dois ou mais avaliadores quanto a uma ou mais características em estudo. Por exemplo, pode-se investigar o grau de concordância entre dois avaliadores quanto à classificação dos n indivíduos entre portadores ou não-portadores de uma dada doença. O objetivo desta monografia é o estudo de aspectos metodológicos relativos à concordância entre avaliadores. Em particular, apresentamos um breve estudo sobre algumas deficiências da Estatística Kappa proposta por Cohen et al. (1960).

Segundo Gwet (2002), as duas estatísticas principais utilizadas para avaliar a intensidade de concordância entre avaliadores são a π-statistic (deveria ser lida como "estatística-pi") sugerida por Scott (1955) e a Kappa Statistic sugerida por Cohen et al. (1960). De acordo com Gwet, pesquisadores comumente usam a π-statistic e, erroneamente, referem-se a ela como Kappa. Dessa forma, faz-se necessário distinguir essas duas estatísticas, que produzirão resultados similares na maioria das vezes – isto será especialmente verdade quando a concordância entre avaliadores é razoavelmente alta.

Consoante com Gwet (2002), a π-statistic e a estatística Kappa serão definidas e alguns exemplos simples serão apresentados para mostrar a inadequação do Kappa para avaliação da extensão de concordância entre 2 avaliadores. Este exemplo será seguido por uma discussão sobre as causas do problema e mais adiante, será dada uma alternativa para possível estatística que produzirá resultados mais recomendados.

2 Objetivos

No presente capítulo apresentamos o objetivo geral e os objetivos específicos da proposta desta Monografia. Tais objetivos nortearão todo o andamento deste trabalho.

2.1 Objetivo Geral

Fornecer detalhes para a discussão sobre as deficiências da Estatística Kappa na concordância entre avaliadores e apontar medidas alternativas. Esta Monografia é baseada no trabalho de Gwet (2002) e o objetivo dela é implementar computacionalmente os métodos discutidos nesse artigo e expandir algumas explanações do mesmo.

2.2 Objetivo Específicos

- Estudar medidas de concordância
- Apresentar exemplos de configurações de dados em que a estatística Kappa apresenta resultados inconsistentes com a evidência amostral
- Comparar a estatística Kappa com outras medidas

3 Metodologia

3.1 Medidas de Concordância

Shoukri (2004) fornece um exemplo de um estudo feito por Westlund and Kurland (1953), em que dois neurologistas revisavam o mesmo conjunto de seleção de atas médicas de potenciais pacientes com esclerose múltipla e classificaram cada um dos indivíduos envolvidos em uma de quatro categorias: 11 (classificação de um indivíduo como portador de esclerose múltipla por ambos os neurologistas), 10 ou 01 (classificação como portador de doença por apenas um dos neurologistas) ou 00 (ambos os neurologistas classificaram um individuo como não portador da doença). O propósito deste estudo é o de determinar a intensidade em que os neurologistas concordaram com os diagnósticos de esclerose múltipla baseado na revisão da ficha médica. A Tabela 1 sumariza os resultados hipotéticos de tal estudo.

Tabela 1 – Uma tabela básica 2×2

		Avaliador 1 (X_1)		
		(1) **Doente**	(0) **Não-Doente**	
Avaliador 2 (X_2)	(1) Doente	n_{11}	n_{10}	$n_{1.}$
	(0) Não-Doente	n_{01}	n_{00}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

Os resultados de um estudo de ensaio e reensaio são usualmente resumidos numa tabela $C \times C$, onde C é o número de categorias nas quais os sujeitos podem ser classificados. A situação mais simples é quando tem-se uma classificação dicotômica (isto é, doença ou não-doença, presença ou ausência, exposto ou não-exposto, etc) resultando uma tabela 2×2. As categorias de resposta são disjuntas (isto é, não há sobreposição).

Uma forma direta de mensurar a intensidade da concordância é computando a quantidade $P_0 = (n_{11} + n_{00})/n$, que é a proporção de concordância total. Este índice é chamado coeficiente de "concordância-simples" (ou concordância global) e a variância estimada desse coeficiente é dada por $Var(P_0) = (P_0(1 - P_0))/n, 0 \leq P_0 \leq 1$. Contudo, de acordo com Shoukri (2004), este índice tem sido duramente criticado, visto que existe uma diferença fundamental entre concordância e associação. Entretanto, para tabelas 2×2, concordância e associação tornam-se indistinguíveis sob certas condições. Essa falha será desenvolvida nas seções a seguir.

Em conformidade com Gwet (2002), duas estatísticas são utilizadas frequentemente para prática de avaliação da extensão da concordância entre avaliadores. Temos a π-statistic (deveria ser lida como "estatística-pi") sugerida por Scott (1955) e a Kappa Statistic sugerida por Cohen et al. (1960). Os pesquisadores comumente usam a π-statistic e erroneamente referem-se a ela como Kappa. É necessário distinguir essas duas estatísticas, que produzirão resultados similares na maioria das vezes (especialmente verdade quando a concordância entre avaliadores é razoavelmente alta).

Depois da π-statistic e a estatística Kappa serem definidas, alguns exemplos serão apresentado para mostrar as deficiências do Kappa para avaliação da extensão da concordância entre 2 avaliadores. Exemplos estes que serão seguidos por uma discussão sobre as causas do problema.

3.2 Estatística π de Scott

Scott (1955) sugeriu computar a extensão da concordância entre os avaliadores 1 e 2 (X_1 e X_2) usando a π-statistic, a qual é definida a seguir:

$$\hat{\pi} = \frac{P_0 - e(\pi)}{1 - e(\pi)}, \tag{3.1}$$

em que $P_0 = (n_{11} + n_{00})/n$ é a proporção de concordância total e $e(\pi)$ é dado por:

$$e(\pi) = \left(\frac{\frac{(n_{.1}+n_{1.})}{2}}{n} \right)^2 + \left(\frac{\frac{(n_{.2}+n_{2.})}{2}}{n} \right)^2 . \tag{3.2}$$

Deve-se observar que o $e(\pi)$ indica a propensão de dos avaliadores concordarem ao acaso, sem terem a mesma avaliação do sujeito.

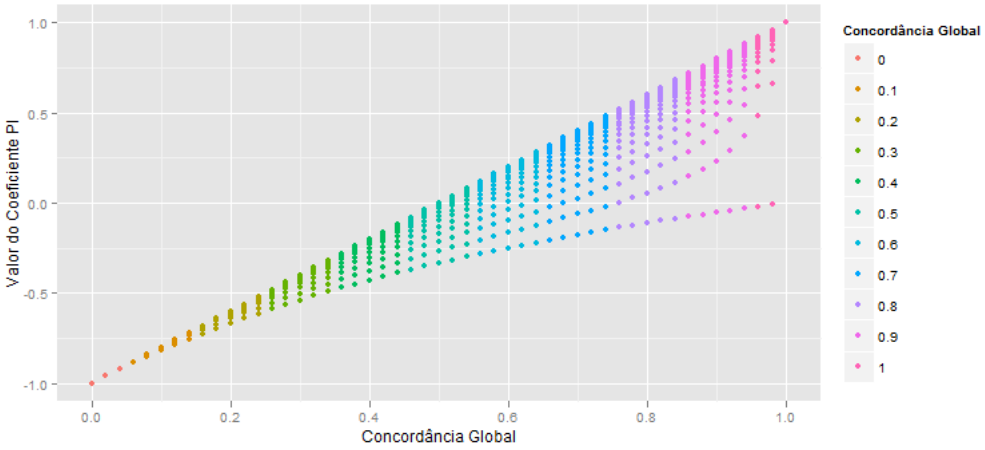
O componente P_0 da equação 3.1 somente envolve os sujeitos que ambos os avaliadores classificaram na mesma categoria. P_0 pode ser utilizada como uma medida ingênua de concordância. Entretanto, há razões para acreditar que os avaliadores 1 e 2 podem classificar alguns sujeitos, na mesma categoria, não pelas mesmas razões. Sujeitos classificados na mesma categoria por diferentes razões correspondem a **concordância ao acaso**. De modo a corrigir a medida de concordância para esse tipo de situação, Gwet (2001) discute extensivamente sobre a motivação da forma da equação 3.1 e explica porque a estatística desse tipo fornece o desejado ajuste.

Infelizmente, identificar os sujeitos que levaram à concordância ao acaso é impossível. Portanto, é necessário calcular a chance de ocorrência dessa concordância devido à sorte. Este aspecto do problema conduziu a controversas propostas por pesquisadores e bioestatísticos. Scott (1955) recomenda a equação (3.2) como uma medida de probabilidade de concordância devida ao acaso. O primeiro termo da equação (3.2) representa a estimação da chance que os avaliadores X_1 e X_2 classifiquem, independentemente, o sujeito

na categoria 0. O segundo termo, por outro lado, estima a probabilidade de ambos os avaliadores classificarem, independentemente, o sujeito na categoria 1.

Ajustando um gráfico em que compara-se a concordância global e o valor do coeficiente Pi proposto por Scott (1955), pode-se observar que, mesmo para valores em que tem-se uma alta concordância global, Pi pode apresentar valores muito próximos de zero para alguns pontos.

Figura 1 – Concordância global em comparação com os possíveis valores da estatística pi



3.3 Estatística Kappa de Cohen

3.3.1 Kappa de Cohen: Medida de Concordância com Chance-Corrigida

Cohen et al. (1960) propôs a estatística Kappa como uma medida de concordância corrigida, descontando a proporção observada de concordância pelo nível de concordância esperada, quando observado a distribuição marginal das respostas dos avaliadores sob a hipótese que os avaliadores fazem avaliações que são estatisticamente independentes. Considere o caso de dois avaliadores, que classificam n sujeitos em uma ou C (onde $C = 2$) categorias nominais mutuamente exclusivas e completas (ou seja, não há sobreposição). Esses avaliadores classificam os sujeitos independentemente. Visto que a concordância observada é

$$P_0 = \frac{n_{11} + n_{00}}{n},$$

o coeficiente Kappa proposto por Cohen et al. (1960) é

$$\hat{\kappa} = \frac{P_0 - e(\kappa)}{1 - e(\kappa)}, \tag{3.3}$$

onde $e(\kappa)$, que é a concordância devida apenas ao acaso, é

$$e(\kappa) = \left(\frac{n_{.1}}{n}\right)\left(\frac{n_{1.}}{n}\right) + \left(\frac{n_{.2}}{n}\right)\left(\frac{n_{2.}}{n}\right). \tag{3.4}$$

Fleiss et al. (1969) fornecem uma aproximação assintótica para a variância estimada de $\hat{\kappa}$, dada como:

$$\begin{aligned}\hat{V}ar(\kappa) = \frac{1}{n(1-e(\kappa))^2}\Bigg(&\sum_{i=2}^{2}\hat{P}_{ii}\{1-(\hat{P}_{i.}+\hat{P}_{.i})(1-\hat{\kappa})\}^2 \\ &+(1-\hat{\kappa})^2\sum_{i\neq j}^{2}\hat{P}_{ij}(\hat{P}_{i.}+\hat{P}_{.j})^2-\{\hat{\kappa}-e(\kappa)(1-\hat{\kappa})\}^2\Bigg).\end{aligned} \tag{3.5}$$

O primeiro termo da equação (3.4) estima a chance de ambos os avaliadores classificarem o sujeito na categoria 00 independentemente. O segundo termo, por sua vez, estima a probabilidade da classificação independente dos sujeitos na categoria 11. Diferente da equação (3.2) de Scott, os termos da equação (3.4) são obtidos através da multiplicação individuais das classificações de cada avaliador. No anexo deste trabalho, temos uma função programada por Gwet (2002) para a estatística kappa no software R (ver anexo 4.1.1).

Cohen et al. (1960) criticou a abordagem de Scott (1955) para o cálculo da probabilidade de concordância ao acaso porque esta combina os dados de classificação dos avaliadores 1 e 2. De fato, o primeiro termo da equação (3.2) é obtido pela média das proporções de sujeitos classificados na categoria 00 por ambos avaliadores, e somado com a média elevada ao quadrado. Essa abordagem elimina qualquer diferença que possa existir na classificação padrão de ambos avaliadores.

Apesar das diferenças entre as estatísticas $\hat{\pi}$ e $\hat{\kappa}$ no modo como a probabilidade de concordância ao acaso é estimada, os dois métodos geralmente apresentam resultados similares. Contudo, a estatística π generaliza o caso de múltiplos avaliadores e múltiplos itens de categoria de resposta mais naturalmente que a estatística κ. Fleiss (1971) propôs a generalização da estatística π de Scott, que é frequentemente referido como estatística Kappa.

3.3.2 Limitações da Estatística Kappa

Segundo Shoukri (2004), apesar da popularidade da estatística Kappa como medida de concordância entre avaliadores, esta apresenta várias limitações e desvantagens. Por exemplo, esse índice depende fortemente da prevalência real da condição que está sendo estudada. Por exemplo, na avaliação dos marcadores de diagnósticos, é sabido que certos testes clínicos, apesar de terem alta sensibilidade e especificidade, podem apresentar baixa capacidade preditiva. Analogamente, dois avaliadores que podem apresentar uma alta concordância, podem produzir um baixo valor para a estatística Kappa. Isto foi esclarecido por Kraemer (1979) que mostrou como a prevalência da condição em estudo altera os resultados do valor da estatística Kappa.

Feinstein and Cicchetti (1990) percebem outro problema ao notar que, na ausência de um "padrão ouro" para servir de referência, não é possível calcular a sensibilidade e especificidade. Por essa razão, a variabilidade entre os avaliadores é melhor descrita pelo coeficiente Kappa. Durante essas considerações, os investigadores encontram um impressionante paradoxo: muitas vezes apesar do relativamente alto valor para a proporção bruta de concordância entre avaliadores (concordância global), o valor de Kappa pode ser relativamente baixo. Se inspecionarmos a expressão da concordância corrigida,

$$\hat{\kappa} = \frac{P_0 - e(\kappa)}{1 - e(\kappa)}$$

é evidente que, dado um valor fixo de P_0, obtém-se um alto valor de Kappa quando $e(\kappa)$ é o menor possível. Considere a seguinte tabela:

P_0	$e(\kappa)$	Kappa
0.85	0.50	0.70
0.85	0.78	0.32

Então, para diferentes valores de $e(\kappa)$, o Kappa, para idênticos valores de P_0, pode ser mais que duas vezes mais alto, quando em comparação com outro caso.

Feinstein and Cicchetti (1990) fornecem a seguinte explicação: o baixo valor de Kappa, apesar do alto valor para P_0, ocorrerá somente quando os totais marginais são drasticamente e simetricamente desbalanceados (texto original: *highly symmetrically unbalanced*). Essa situação ocorre quando $n_{1.}$ (total da primeira coluna) é muito diferente de $n_{2.}$ (total da segunda coluna), ou quando o $n_{.1}$ é muito diferente de $n_{.2}$. Um equilíbrio perfeito ocorre quando $n_{1.} = n_{2.}$ ou quando $n_{.1} = n_{.2}$. Como exemplo, considere a seguinte tabela:

Tabela 2 – Exemplo de tabela simetricamente desbalanceada

		Avaliador 1			
		Sim	Não	Total	
Avaliador 2	Sim	40	5	45	$n_{1.}$
	Não	3	2	5	
	Total	43	7	50	$n_{2.}$
		$n_{.1}$	$n_{.2}$		n

Note que $P_0 = 0.84$ indica um alto nível de concordância observada. Contudo,

$$e(\kappa) = \left(\frac{43}{50}\right)\left(\frac{45}{50}\right) + \left(\frac{7}{50}\right)\left(\frac{5}{50}\right) = 0.79$$

e

$$\hat{\kappa} = \frac{(.84 - .79)}{(1 - .79)} = 0.24,$$

indicando, contrariamente a P_0, baixíssima concordância. Como pode-se observar da tabela 2, esse paradoxo é causado por conta da acentuada diferença entre $n_{.1} = 43$ e $n_{.2} = 7$, ou por causa da marcada diferença entre $n_{1.} = 45$, e $n_{2.} = 5$.

Tabela 3 – Exemplo de Tabela Assimetricamente Desbalanceada

		Avaliador 1			
		Sim	Não	Total	
	Sim	21	6	27	$n_{1.}$
Avaliador 2	Não	2	21	23	
	Total	23	27	50	$n_{2.}$
		$n_{.1}$	$n_{.2}$		n

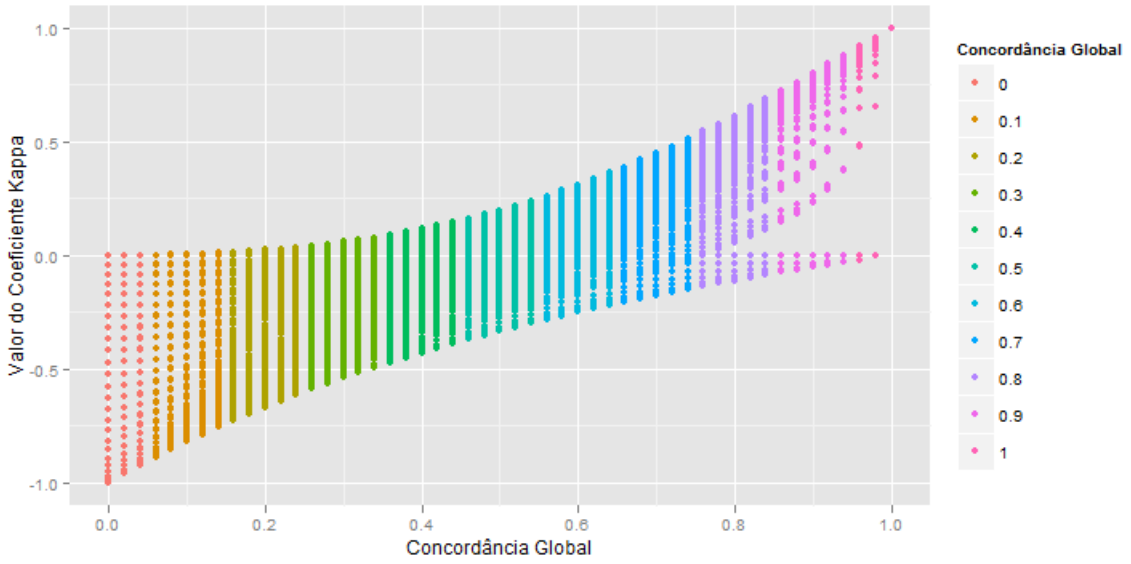
Um segundo paradoxo ocorre quando totais marginais desbalanceados produzem altos valores de Kappa comparado com totais mais balanceados. Essa situação ocorre quando $n_{.1}$ é muito maior que $n_{.2}$, enquanto $n_{1.}$ é muito menor que $n_{2.}$, ou vice-versa. A mesma situação, que produz "assimetria desbalanceada marginal", é ilustrada na tabela 3. Aqui, $P_0 = 0.84$, $e(\kappa) = 0.50$ e $\kappa = 0.68$, que é um valor muito maior que o valor obtido quando a estatística Kappa é oriunda de uma tabela simetricamente desbalanceada.

De forma a ilustrar, graficamente, essa diferença, a figura 2 apresenta os valores obtidos no coeficiente de concordância da estatística Kappa e com concordância global (também chamada, nesse trabalho, de concordância observada). O gráfico foi produzido ao simular todas as alocações possíveis para uma tabela com $n = 50$ (que gera 23426 possíveis tabelas diferentes). O código completo em R que desenvolvemos para finalidade está descrito no anexo (4.1.4.1), na página 25 desta Monografia.

```
# Para gerar uma lista como todas as matrizes possiveis:
f.matriz <- function(n){
  lista <- list()
  contador <- 1
  for (i in 0:n){
    for (j in 0:(n-i)){
      for (k in 0:(n-i-j)){
        vetor <- c(i, j, k, n-i-j-k)
        lista[[contador]] <- vetor
        contador <- contador + 1
      }
    }
  }
  return(lista)
}
```

```
# Exemplo: n=1
f.matriz(1)
##   [[1]]
##   [1] 0 0 0 1

##   [[2]]
##   [1] 0 0 1 0

##   [[3]]
##   [1] 0 1 0 0

##   [[4]]
##   [1] 1 0 0 0

# Obter todas as possibilidades para n fixo
n<-50
a<-f.matriz(n)
n.com<-length(a)
n.com
## [1] 23426
```

Figura 2 – Concordância global em comparação com o valor da estatística kappa



Repare que, mesmo em pontos onde tem-se uma alta concordância global, pode-se obter baixos valores da estatística Kappa.

Reproduzindo, agora, os exemplos de Gwet (2002), considere 2 experimentos denominados E1 e E2. Para cada um dos experimentos, os avaliadores A e B têm que classificar 100 sujeitos em 2 categorias de resposta, nomeadas de "1" e "2". As tabelas 4 e 5 a seguir descrevem os resultados dos experimentos E1 e E2 respectivamente.

Tabela 4 – Resultados do experimento E1

		Avaliador A		
		"1"	**"2"**	**Total**
Avaliador B	**"1"**	40	9	49
	"2"	6	45	51
	Total	46	54	100

Tabela 5 – Resultados do experimento E2

		Avaliador A		
		"1"	**"2"**	**Total**
Avaliador B	**"1"**	80	10	90
	"2"	5	5	10
	Total	85	15	100

Pode-se observar que, em ambos os experimentos, os avaliadores A e B concordaram sobre a classificação de 85 sujeitos. Portanto, a concordância global P_0 é igual a 0.85 em ambos os experimentos. É sensato esperar uma alta concordância inter-avaliadores nos avaliadores em ambas as situações. Infelizmente, nem a estatística π de Scott, nem a estatística κ de Cohen fornecem estimativas coerentes de concordância em ambos os experimentos.

Com a estatística de Scott para os dois experimentos obteve-se os seguintes valores:

- Para o experimento E1:

  Probabilidade de Concordância ao acaso $e(\pi)$ é dada por:

$$e(\pi) = \left(\frac{\frac{(46+49)}{2}}{100}\right)^2 + \left(\frac{\frac{(51+54)}{2}}{100}\right)^2 = 0.50125,$$

  que conduz à estatística π :

$$\hat{\pi} = \frac{0.85 - 0.50125}{1 - 0.50125} = 0.6993.$$

- Para o experimento E2:

  Probabilidade de Concordância ao acaso $e(\pi)$ é dada por:

$$e(\pi) = \left(\frac{\frac{(90+85)}{2}}{100}\right)^2 + \left(\frac{\frac{(10+15)}{2}}{100}\right)^2 = 0.78125,$$

que conduz a estatística π :

$$\hat{\pi} = \frac{0.85 - 0.78125}{1 - 0.78125} = 0.3143.$$

A estatística Kappa de Cohen, por outro lado, resulta nos seguintes valores:

- Para o experimento E1:

  Probabilidade de concordância ao acaso, $e(\kappa)$, é dada por:

  $$e(\kappa) = \left(\frac{49}{100}\right)\left(\frac{46}{100}\right) + \left(\frac{51}{100}\right)\left(\frac{54}{100}\right) = 0.5008,$$

  que conduz a estatística κ :

  $$\hat{\kappa} = \frac{(0.85 - 0.5008)}{(1 - 0.5008)} = 0.6994.$$

- Para o experimento E2:

  Probabilidade de concordância ao acaso, $e(\kappa)$, dada por :

  $$e(\kappa) = \left(\frac{90}{100}\right)\left(\frac{85}{100}\right) + \left(\frac{10}{100}\right)\left(\frac{15}{100}\right) = 0.78,$$

  que conduz a estatística κ :

  $$\hat{\kappa} = \frac{(0.85 - 0.78)}{(1 - 0.78)} = 0.318.$$

Pode-se também obter esses valores através do software R, com as funções programadas por Gwet (2002) que estão em anexo. Através delas, obtém-se:

```
> scott2.table(E1)
Scott's Pi Coefficient
==========================
Percent agreement: 0.85 Percent chance agreement: 0.50125
Scott coefficient: 0.6992481 Standard error: 0.07158374
95 % Confidence Interval: ( 0.5572104 , 0.8412858 )
P-value:  4.440892e-16

> kappa2.table(E1)
Cohen's Kappa Coefficient
==========================
Percent agreement: 0.85 Percent chance agreement: 0.5008
```

```
Kappa coefficient: 0.6995192 Standard error: 0.0713936
95 % Confidence Interval: ( 0.5578588 , 0.8411796 )
P-value:  2.220446e-16

> scott2.table(E2)
Scott's Pi Coefficient
======================
Percent agreement: 0.85 Percent chance agreement: 0.78125
Scott coefficient: 0.3142857 Standard error: 0.1354768
95 % Confidence Interval: ( 0.04547043 , 0.583101 )
P-value:  0.02240268

> kappa2.table(E2)
Cohen's Kappa Coefficient
=========================
Percent agreement: 0.85 Percent chance agreement: 0.78
Kappa coefficient: 0.3181818 Standard error: 0.1334565
95 % Confidence Interval: ( 0.05337513 , 0.5829885 )
P-value:  0.01902393
```

A estatística κ e π surpreendentemente indicam um baixo nível de concordância entre os avaliadores A e B no experimento E2. De fato, é difícil explicar porque os avaliadores teriam um alto nível de concordância no experimento E1 e uma concordância razoavelmente baixa no experimento E2. Esse paradoxo levou muitos autores a concluir que a estatística Kappa é dramaticamente afetada pela traço de prevalência da população em consideração. Outros cientistas recomendam o teste de homogeneidades marginais para determinar a adequação da estatística Kappa.

Seguindo as explicações de Shoukri (2004), tem-se que, em resumo, desequilíbrios na distribuição dos totais marginais podem produzir dois tipos de paradoxos quando a variabilidade nas classificações binárias de 2 avaliadores é expressa através do coeficiente Kappa. O Kappa pode ser baixo, apesar dos altos valores de P_0. Pode ocorrer de assumir valores altos, ao invés de reduzidos, por conta da assimetria dos totais marginais. Apesar desses problemas paradoxais, o Kappa é um índice versátil no cálculo de concordância nominal entre dois avaliadores. Contudo, Gwet (2002) acredita que há sérias falhas conceituais em ambas as estatísticas (κ e π) tornando-as não confiáveis.

3.3.3 Origens da inadequação das Estatísticas PI e KAPPA

Gwet (2002) afirma que a forma geral das estatísticas Kappa e π, como funções de P_0 e e (tanto $e(\pi)$ como $e(\kappa)$), são apropriadas para a correção da propensão à concordância

ao acaso. Entretanto, é a expressão usada no cálculo da probabilidade de concordância devida ao acaso que é inapropriada.

De modo a obter uma expressão adequada para a probabilidade de concordância ao acaso, é necessário definir o que é concordância ao acaso e explicar as circunstâncias na qual esta ocorre. Qualquer concordância entre 2 avaliadores A e B, pode ser considerada com a concordância ao acaso, se o avaliador realiza uma classificação aleatória (isto é, classifica o sujeito sem ser guiado pelas características do mesmo) e ambos os avaliadores concordam em suas classificações. Se a classificação é aleatória, Gwet (2002) afirma que é possível demonstrar que a concordância pode ocorrer com uma probabilidade fixada de 0.5. Simulações que foram conduzidas por Gwet (2001) também tendem a confirmar esse fato. Segue que o valor razoável para a probabilidade de concordância ao acaso não deveria exceder a 0.5.

Figura 3 – Função da probabilidade de classificação dos avaliadores para estatística Pi

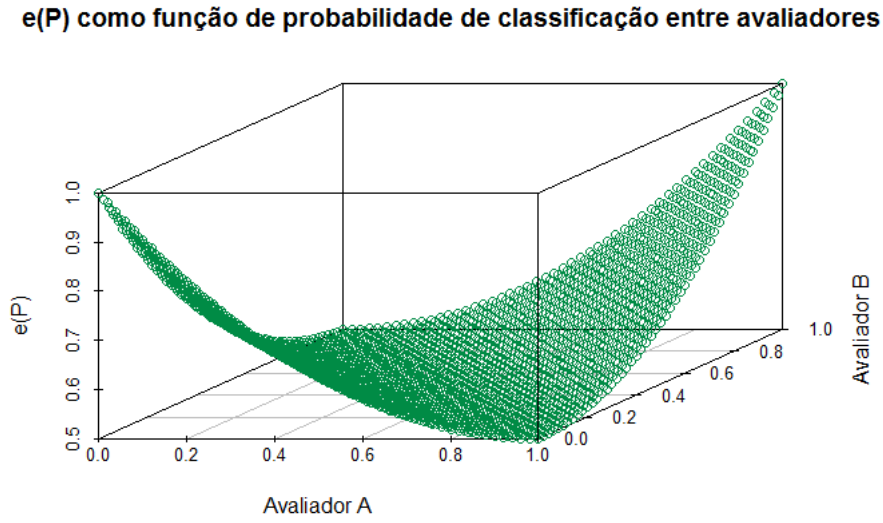
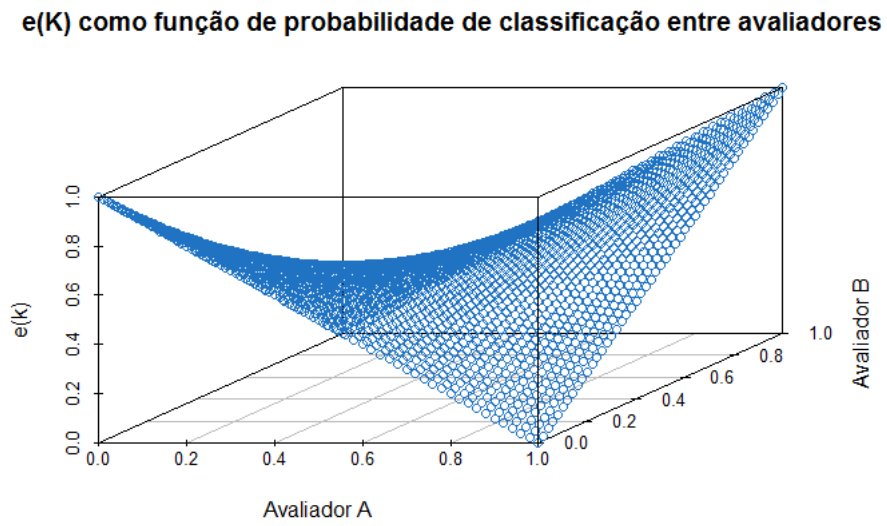


Figura 4 – Função da probabilidade de classificação dos avaliadores para estatística Kappa



As Figuras 3 e 4 mostram um gráfico da probabilidade de chance de concordância para as estatísticas Pi e Kappa, como a função das probabilidades de classificação marginal dos avaliadores na categoria 1. A probabilidade marginal é $P1_A = n_{.1}/n$ para o avaliador A e $P1_B = n_{1.}/n$ para o avaliador 2. Segue da Figura 3 que a probabilidade de concordância ao acaso $e(\pi)$ para a estatística Pi varia de 0.5 a 1. Esta propriedade contradiz, seriamente, a constatação do parágrafo anterior, o qual sugere que um valor razoável para a probabilidade de chance de concordância não deveria exceder 0.5. De fato, todos os valores de $e(\pi)$ ultrapassam 0.5. É difícil imaginar uma circunstância onde 2 avaliadores concordam, ao acaso, com probabilidade de 1. Ademais, a Figura 3 também sugere que se $n_{.1} = 0$ e $n_{1.} = 0$, então os avaliadores A e B concordarão ao acaso com probabilidade 1.

A Figura 4 mostra o gráfico de $e(\kappa)$ como uma função da classificação marginal das probabilidades $P1_A$ e $P1_B$. Essa Figura mostra que a probabilidade de concordância ao acaso $e(\kappa)$ pode assumir qualquer valor entre 0 e 1, que viola a condição de estar acima do limite superior de 0.5. O valor de $e(\kappa)$ é geralmente menor que 0.5 se a soma das probabilidades de classificação marginais forem razoavelmente próximas de 1. A Figura 4 também indica que as piores situações da estatística Kappa ocorrem quando ambas as probabilidades de classificação marginais $P1_A$ e $P1_B$ são muito pequenas. Há situações onde a curva de $e(\kappa)$ fica mais perto do valor máximo 1.

Para construir semelhantes as Figuras 3 e 4, utilizou-se um pacote chamado "Rcommander", que é um pacote complementar ao software estatístico R que possibilita a interação com o gráfico. Assim, para a melhor visualização dos gráficos através da rotação das imagens num gráfico tridimensional, pode-se executar os comandos em anexo (vide página 27, em 4.1.4.3) a fim de explorar melhor as superfícies.

3.3.4 Alternativa de correção para cálculo do acaso na estatística Kappa

Uma simples alternativa mais sensata estatisticamente poderia ser usada para estimar a extensão da concordância entre os avaliadores. Na seção anterior, considerou-se a concordância ao acaso como sendo a ocorrência simultânea ou classificação aleatória (por um dos avaliadores) e a concordância entre avaliadores. Portanto, calcula-se a probabilidade de concordância ao acaso multiplicando a *propensão da ocorrência de um avaliador fazer uma avaliação ao acaso* pela probabilidade de haver concordância sob a suposição de que a avaliação foi feita ao acaso, consoante a Gwet (2002).

A probabilidade da ocorrência de concordância ao acaso é 0.5. Isto é, se pelo menos um dos dois avaliadores faz classificações ao acaso, os avaliadores concordarão em até 50% das vezes. Gwet (2001) provou tal afirmação matematicamente e por meio de simulação. A propensão à classificações ao acaso é definida em termos da proporção da máxima variância observada no experimento. Maiores detalhes estão descritos em Gwet (2001).

Seja $e(\gamma)$ a nova probabilidade de ocorrência de concordância ao acaso. Tem-se que:

$$e(\gamma) = 2P_a(1 - P_a) \tag{3.6}$$

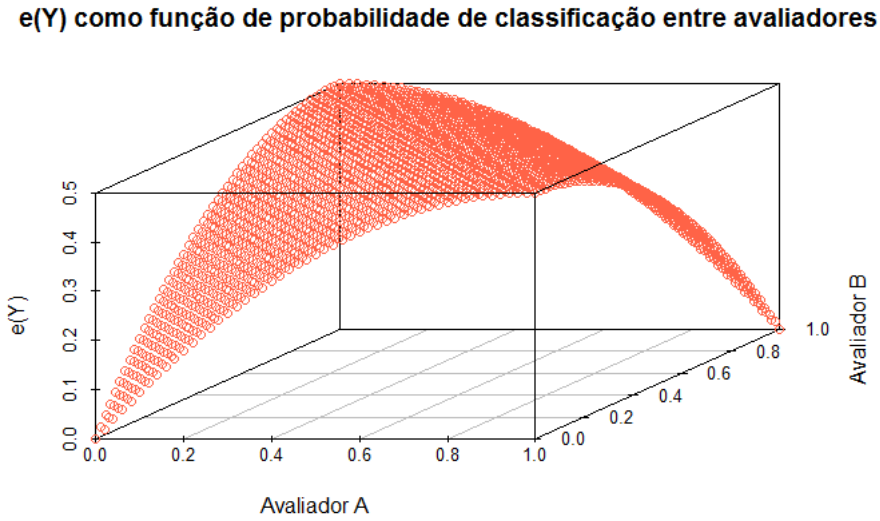
onde $P_a = \frac{(n_{1.}+n_{.1})}{2}/n$ representa a chance aproximada de um avaliador (A ou B) classifique um individuo na categoria 1. A estatística alternativa, a qual é referida como estatística AC1 de Gwet (2001) é dada por:

$$AC1 = \frac{P_0 - e(\gamma)}{1 - e(\gamma)} \tag{3.7}$$

onde $P_0 = (n_{11} + n_{22})/n$ e $e(\gamma)$ é dado pela equação (3.6).

De modo a tornar possível a comparação, a Figura 5 retrata a distribuição de $e(\gamma)$ como uma função das probabilidades de classificação de ambos avaliadores.

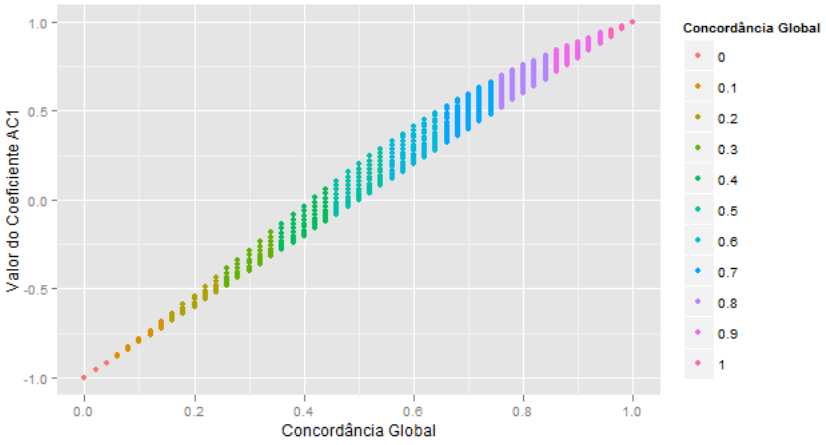
Figura 5 – Função da probabilidade de classificação dos avaliadores para estatistica AC1



Segue da Figura 5 que a probabilidade de concordância ao acaso, $e(\gamma)$, sempre varia entre 0 e 0.5. Essa probabilidade é próxima ao valor máximo de 0.5 se a soma das probabilidades marginais é por volta de 1, e decresce, a medida que a soma se afasta de 1.

Ao comparar o gráfico de probabilidade de concordância ao acaso com os possíveis valores obtidos para a estatística de AC1, observa-se que há uma forte tendência linear descrita entre AC1 e P_0, com pouca variabilidade dos pontos em torno de uma reta imaginária. Comparando a Figura 6 com a Figura 2 pode-se avaliar, mais precisamente, a superioridade da estatística AC1 em relação a Kappa.

Figura 6 – Concordância global em comparação com o valor da estatística AC1



- Para o experimento E1 da seção anterior, a probabilidade de concordância ao acaso de $e(\gamma)$ é dada por:

$$e(\gamma) = 2\left(\frac{(46+49)}{(2\times 100)}\right)\left(\frac{1-(46+49)}{(2\times 100)}\right) = 0.49875,$$

e conduz a estatística AC1 :

$$AC1 = \frac{0.85 - 0.49875}{1 - 0.49875} = 0.7008.$$

Essa concordância entre os avaliadores é bem semelhante aos valores obtidos com as estatísticas Pi e Kappa (0.6993 e 0.6994, respectivamente).

- Para o experimento E2 por sua vez, a probabilidade de concordância ao acaso de $e(\gamma)$ é dada por:

$$e(\gamma) = 2\left(\frac{(90+85)}{(2\times 100)}\right)\left(\frac{1-(90+85)}{(2\times 100)}\right) = 0.21875,$$

e conduz a estatística AC1:

$$AC1 = \frac{0.85 - 0.21875}{1 - 0.21875} = 0.808$$

Para esse experimento, a concordância entre avaliadores que fora estimada foi de 0.3143 e 0.318 para as estatísticas Pi e Kappa respectivamente. Parece claramente que a estatística AC1 fornece uma estimação que é mais consistente com os resultados do experimento E2.

Pode-se, também, obter esses valores através do software R, com a função programada no anexo (4.1.2). Assim, tem-se para os experimentos E1 e E2:

```
> gwet.ac1.table(E1)

Gwet's AC1/AC2 Coefficient
==========================
Percent agreement: 0.85 Percent chance agreement: 0.49875
AC1/AC2 coefficient: 0.7007481 Standard error: 0.0713518
95 % Confidence Interval: ( 0.5591707 , 0.8423256 )
P-value:  2.220446e-16

> gwet.ac1.table(E2)
Gwet's AC1/AC2 Coefficient
==========================
Percent agreement: 0.85 Percent chance agreement: 0.21875
AC1/AC2 coefficient: 0.808 Standard error: 0.05212942
95 % Confidence Interval: ( 0.7045639 , 0.9114361 )
P-value:  0
```

Utilizando a tabela 6, condensa-se as informações dos experimentos que foram detalhados anteriormente. Dessa forma, com esses experimentos abordados, nota-se que a estatística AC1 avalia a concordância entre avaliadores de forma mais robusta, visto que não altera-se indevidamente por conta da disposição da tabela.

Tabela 6 – Comparação de Resultados

	Concordância Global	Estatística de Scott	Estatística de Cohen	Estatística de Gwet
	P_0	π	κ	$AC1$
E1	0.850	0.699	0.699	0.701
E2	0.850	0.314	0.318	0.808

4 Conclusão

O objetivo dessa Monografia foi triplo:

- Provar que as estatísticas Pi e Kappa, largamente utilizadas, podem ser enganosas em muitos casos, especialmente quando a soma das probabilidades marginais é muito diferente de 1;
- Introduzir uma estatística com probabilidade de concordância ao acaso como alternativa mais robusta que consistentemente rende resultados mais Sensatos;
- Fornecer detalhes sobre as deficiências da estatística Kappa e implementar computacionalmente os métodos desenvolvidos por Gwet (2002);

Ao observar o comportamento imprevisível das estatísticas Pi e Kappa, percebeu-se que ele ocorre devido a um método equivocado no cálculo da probabilidade de concordância ao acaso. Isto tem infelizmente levado a alguns pesquisadores questionarem muito o mérito da estatística Kappa na correlação e da concordância entre avaliadores ao acaso. A estatística AC1 fornece uma abordagem mais coerente e robusta.

Referências

Bussab, W. d. O. and Morettin, P. A. (2010). *Estatística básica*. Saraiva.

Cohen, J. et al. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fleiss, J. L., Cohen, J., and Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323.

Gwet, K. (2001). Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD: STATAXIS Publishing Company*.

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment*, 1(6):1–6.

Kraemer, H. C. (1979). Ramifications of a population model forκ as a coefficient of reliability. *Psychometrika*, 44(4):461–472.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*.

Shoukri, M. (2004). *Measures of interoberver agreement*. Chapman & Hall/CRC.

Westlund, K. B. and Kurland, L. T. (1953). Studies on multiple sclerosis in winnipeg, manitoba, and new orleans, louisiana, ii. a controlled investigation of factors in the life history of the winnipeg patients. *American Journal of Epidemiology*, 57(3):397–407.

Anexo

4.1 Programação em R

4.1.1 Estatística Kappa programada por Gwet

```
kappa2.table <- function(ratings,weights=diag(ncol(ratings)),
conflev=0.95,N=Inf,print=TRUE){
  if(dim(ratings)[1] != dim(ratings)[2]){
    stop('The contingency table should have the
    same number of rows and columns!')
  }
  n <- sum(ratings) # number of subjects
  f <- n/N # final population correction
  q <- ncol(ratings) # number of categories
  pa <- sum(weights * ratings/n) # percent agreement

  pk. <- (ratings%*%rep(1,q))/n
  p.l <- t((t(rep(1,q))%*%ratings)/n)
  pe <- sum(weights*(pk.%*%t(p.l)))
  kappa <- (pa - pe)/(1 - pe) # weighted kappa

  # 2 raters special case variance

  pkl <- ratings/n
  pb.k <- weights %*% p.l
  pbl. <- t(weights) %*% pk.
  sum1 <- 0
  for(k in 1:q){
    for(l in 1:q){
      sum1 <- sum1 + pkl[k,l]* (weights[k,l]-
      (1-kappa)*(pb.k[k] + pbl.[l]))^2
    }
  }
  var.kappa <- ((1-f)/(n*(1-pe)^2)) *
  (sum1 - (pa-2*(1-kappa)*pe)^2)
  stderr <- sqrt(var.kappa) # kappa standard error
  p.value <- 2*(1-pt(kappa/stderr,n-1))
```

```
  lcb <- kappa - stderr*qt(1-(1-conflev)/2,n-1)
  # lower confidence bound
  ucb <- min(1,kappa + stderr*qt(1-(1-conflev)/2,n-1))
  # upper confidence bound
  if(print==TRUE){
    cat("Cohen's Kappa Coefficient\n")
    cat('========================\n')
cat('Percent agreement:',pa,'Percent chance agreement:',pe,'\n')
cat('Kappa coefficient:',kappa,'Standard error:',stderr,'\n')
cat(conflev*100,'% Confidence Interval: (',lcb,',',ucb,')\n')
cat('P-value: ',p.value,'\n')
  }
  invisible(c(pa,pe,kappa,stderr,p.value))
}
```

4.1.2 Estatística AC1 programada por Gwet

```
gwet.ac1.table <- function(ratings,weights=diag(ncol(ratings)),
conflev=0.95,N=Inf,print=TRUE){
  if(dim(ratings)[1] != dim(ratings)[2]){
    stop('The contingency table should
    have the same number of rows and columns!')
  }
  n <- sum(ratings) # number of subjects
  f <- n/N # final population correction
  q <- ncol(ratings) # number of categories
  pa <- sum(weights * ratings/n) # percent agreement

  pk. <- (ratings%*%rep(1,q))/n
  p.l <- t((t(rep(1,q))%*%ratings)/n)
  pi.k <- (pk.+p.l)/2
  tw <- sum(weights)
  pe <- tw * sum(pi.k *(1-pi.k))/(q*(q-1))
  gwet.ac1 <- (pa - pe)/(1 - pe)
  # gwet's ac1/ac2 coefficint

  # calculation of variance - standard error -
  # confidence interval - p-value
```

```
  pkl <- ratings/n              #p_{kl}
  sum1 <- 0
  for(k in 1:q){
    for(l in 1:q){
      sum1 <- sum1 + pkl[k,l] * (weights[k,l]-2*
      (1-gwet.ac1)*tw*(1-(pi.k[k] + pi.k[l])/2)/(q*(q-1)))^2
    }
  }
  var.gwet <- ((1-f)/(n*(1-pe)^2)) * (sum1 -
  (pa-2*(1-gwet.ac1)*pe)^2)
  stderr <- sqrt(var.gwet)# ac1's standard error
  p.value <- 2*(1-pt(gwet.ac1/stderr,n-1))

  lcb <- gwet.ac1 - stderr*qt(1-(1-conflev)/2,n-1)
  # lower confidence bound
  ucb <- min(1,gwet.ac1 + stderr*qt(1-(1-conflev)/2,n-1))
  # upper confidence bound
  if(print==TRUE){
cat("Gwet's AC1/AC2 Coefficient\n")
cat("==========================\n")
cat('Percent agreement:',pa,'Percent chance agreement:',pe,'\n')
cat('AC1/AC2coefficient:',gwet.ac1,'Standard error:',stderr,'\n')
cat(conflev*100,'% Confidence Interval: (',lcb,',',ucb,')\n')
cat('P-value: ',p.value,'\n')
  }
  invisible(c(pa,pe,gwet.ac1,stderr,p.value))
}
```

4.1.3 Estatística PI programada por Gwet

```
scott2.table <- function(ratings,weights=diag(ncol(ratings)),
conflev=0.95,N=Inf,print=TRUE){
  if(dim(ratings)[1] != dim(ratings)[2]){
    stop('The contingency table should
    have the same number of rows and columns!')
  }
  n <- sum(ratings) # number of subjects
  f <- n/N # final population correction
```

```
  q <- ncol(ratings) # number of categories
  pa <- sum(weights * ratings/n) # percent agreement

  pk. <- (ratings%*%rep(1,q))/n
  p.l <- t((t(rep(1,q))%*%ratings)/n)
  pi.k <- (pk.+p.l)/2
  pe <- sum(weights*(pi.k%*%t(pi.k)))
  scott <- (pa - pe)/(1 - pe) # weighted scott's pi coefficint

  # 2 raters special case variance

  pkl <- ratings/n            #p_{kl}
  pb.k <- weights %*% p.l     #\ov{p}_{+k}
  pbl. <- t(weights) %*% pk.  #\ov{p}_{l+}
  pbk  <- (pb.k + pbl.)/2     #\ov{p}_{k}
  sum1 <- 0
  for(k in 1:q){
    for(l in 1:q){
      sum1 <- sum1 + pkl[k,l] * (weights[k,l]-
      (1-scott)*(pbk[k] + pbk[l]))^2
    }
  }
  var.scott <- ((1-f)/(n*(1-pe)^2)) *
  (sum1 - (pa-2*(1-scott)*pe)^2)
  stderr <- sqrt(var.scott)# Scott's standard error
  p.value <- 2*(1-pt(scott/stderr,n-1))

  lcb <- scott - stderr*qt(1-(1-conflev)/2,n-1)
  # lower confidence bound
  ucb <- min(1,scott + stderr*qt(1-(1-conflev)/2,n-1))
  # upper confidence bound
  if(print==TRUE){
cat("Scott's Pi Coefficient\n")
cat('==========================\n')
cat('Percent agreement:',pa,'Percent chance agreement:',pe,'\n')
cat('Scott coefficient:',scott,'Standard error:',stderr,'\n')
cat(conflev*100,'% Confidence Interval: (',lcb,',',ucb,')\n')
cat('P-value: ',p.value,'\n')
  }
```

```
  invisible(c(pa,pe,scott,stderr,p.value))
}
```

4.1.4 Gráficos

4.1.4.1 Função que fornece valores para os gráficos adotados

```
# Para gerar uma lista como todas as matrizes possiveis:
f.matriz <- function(n){
  lista <- list()
  contador <- 1
  for (i in 0:n){
    for (j in 0:(n-i)){
      for (k in 0:(n-i-j)){
        vetor <- c(i, j, k, n-i-j-k)
        lista[[contador]] <- vetor
        contador <- contador + 1
      }
    }
  }
  return(lista)
}

# Exemplo: n=1
f.matriz(1)
f.matriz(1)[[1]]

# Obter todas as possibilidades para n fixo
n<-50
a<-f.matriz(n)
n.com<-length(a)

# Vetores Associados
a1<-vector()
a2<-vector()
b1<-vector()
b2<-vector()
n<-vector()
p.a1<-vector()
```

```
p.b1<-vector()
p.0<-vector()
eK<-vector()
KP<-vector()
eP<-vector()
PI<-vector()
p.1<-vector()
eY<-vector()
AC1<-vector()

for(i in 1:n.com){
  b<-matrix(a[[i]],2,2)

  a1[i]<-b[1,1]+b[2,1]
  a2[i]<-b[1,2]+b[2,2]
  b1[i]<-b[1,1]+b[1,2]
  b2[i]<-b[2,1]+b[2,2]
  n[i]<-sum(b)

  p.a1[i]<-a1[i]/n[i]   #Probabilidade Marginal de P1A
  p.b1[i]<-b1[i]/n[i]   #Probabilidade Marginal de P1B

  p.0[i]<-(b[1,1]+b[2,2])/n[i]  #Concordancia Global

eK[i]<- (a1[i]/n[i])*(b1[i]/n[i]) + (a2[i]/n[i])*(b2[i]/n[i])
  # Concordancia ao acaso (Kappa)
KP[i]<- (p.0[i] - eK[i])/(1-eK[i])
  # Coeficiente Kappa (Cohen)

eP[i]<- (((a1[i]+b1[i])/2)/n[i])^2 + (((a2[i]+b2[i])/2)/n[i])^2
  # Concordancia ao acaso (PI)
PI[i]<- (p.0[i] - eP[i])/(1-eP[i])
  # Coeficiente PI (Scott)

p.1[i]<-((a1[i]+b1[i])/2)/n[i]
eY[i]<- 2*p.1[i]*(1-p.1[i])   #Concordancia ao acaso (AC1)
AC1[i]<-(p.0[i] - eY[i])/(1-eY[i])  #Coeficiente AC1 (Gwet)
}
```

```
dados<-cbind(a1,a2,b1,b2,n,p.a1,p.b1,p.0,eK,KP,eP,PI,p.1,eY,AC1)
head(dados)
```

4.1.4.2 Gráficos 3D com figuras fixas

```
#=============================================#
# Graficos em 3D com figuras fixas - scatterplot3d  #
#=============================================#
install.packages("scatterplot3d")
require("scatterplot3d")

scatterplot3d(p.a1,p.b1,eK, color=rgb(0.12,0.45,0.76),
main="e(K) como funcao de probabilidade de classificacao
entre avaliadores",
xlab="Avaliador A", ylab="Avaliador B", zlab="e(k)")

scatterplot3d(p.a1,p.b1,eY, color="tomato",
main="e(Y) como funcao de probabilidade de classificacao
entre avaliadores",
xlab="Avaliador A", ylab="Avaliador B", zlab="e(Y)")

scatterplot3d(p.a1,p.b1,eP, color="springgreen4",
main="e(P) como funcao de probabilidade de classificacao
entre avaliadores",
xlab="Avaliador A", ylab="Avaliador B", zlab="e(P)")
```

4.1.4.3 Gráficos 3D com figuras fixas

```
#-------------------------------------------#
# Grafico 3D rotacionado usando o "Rcommander"  #
#-------------------------------------------#

require(Rcmdr)

# Quando aparecer uma janela escrito "Rcommander",
# minimiza-a e continue rodando o que temos abaixo

# Grafico Kappa
scatter3d(p.a1, eK, p.b1, ellipsoid=F, surface=T,
```

```
point.col='yellow', bg='black', sphere.size=1.2,
revolutions=1, axis.col='white', xlab="Avaliador A",
ylab="e(k)", zlab="Avaliador B")

#Grafico PI
scatter3d(p.a1, eP, p.b1, ellipsoid=F, surface=T,
point.col='yellow', bg='black', sphere.size=1.2,
revolutions=1, axis.col='white', xlab="Avaliador A",
ylab="e(p)", zlab="Avaliador B")

#Grafico AC1
scatter3d(p.a1, eY, p.b1, ellipsoid=F, surface=T,
point.col='yellow', bg='black', sphere.size=1.2,
revolutions=1, axis.col='white', xlab="Avaliador A",
ylab="e(Y)", zlab="Avaliador B")
```

4.1.4.4 Gráficos 2D para comparação de valores

```
#-------------------------------------------------------------#
# Comparacao entre a concordancia ao acaso e os coeficientes #
#-------------------------------------------------------------#

require(ggplot2)
dados2<-data.frame(dados)



ggplot(data=dados2, aes(x=p.0 , y=KP,
colour=as.factor(round(p.0,1)))) +
  geom_point() +
  labs(colour = "Concordancia Global") +
  xlab("Concordancia Global") +
  ylab("Valor do Coeficiente Kappa")

ggplot(data=dados2, aes(x=p.0 , y=AC1,
colour=as.factor(round(p.0,1)))) +
  geom_point() +
  labs(colour = "Concordancia Global") +
  xlab("Concordancia Global") +
  ylab("Valor do Coeficiente AC1")
```

```
ggplot(data=dados2, aes(x=p.0 , y=PI,
colour=as.factor(round(p.0,1)))) +
  geom_point() +
  labs(colour = "Concordancia Global") +
  xlab("Concordancia Global") +
  ylab("Valor do Coeficiente PI")
```